Identification of Technology Terms in Patents

Peter Anick, Marc Verhagen and James Pustejovsky
Computer Science Department
Brandeis University
Waltham, MA, United States
peter_anick@yahoo.com, marc@cs.brandeis.edu, jamesp@cs.brandeis.edu

Abstract

Natural language analysis of patents holds promise for the development of tools designed to assist analysts in the monitoring of emerging technologies. One component of such tools is the identification of technology terms. We describe an approach to the discovery of technology terms using supervised machine learning and evaluate its performance on subsets of patents in three languages: English, German, and Chinese.

Keywords: text mining, terminology, patents

1. Introduction

The timely detection of emerging technologies and the monitoring of their worldwide evolution pose daunting challenges for analysts (PICMET, 2012). Not only do these tasks demand constantly expanding domain expertise but the rate of scientific publication is growing fast (Sharma et al., 2002; Larsen and Ins, 2010).

Patent filings represent a leading indicator of the maturation of technologies and their introduction into the marketplace. As semi-structured documents, they offer many opportunities for data mining of natural language content. For example, citations and references to prior art reflect the intellectual development of a technology while the appearance of novel terminology in a cluster of patents suggests the emergence of a new subfield. Previous research on patents has applied natural language processing for the purpose of summarization and clustering (Tseng et al., 2007), infringement analysis (Indukuri et al., 2007), and computer-assisted categorization (Fall et al., 2003). Numerous techniques for the automatic extraction of terms and phrases in support of these tasks have been proposed. However, such efforts have rarely made a distinction between terms that denote technologies and other classes of terms. In this paper, we seek to automate the identification of technology terms within patents in order to make this constantly-growing technical vocabulary available for the construction of higher level analytical tools. This work was developed in the context of an automated system that processes very large collections of patents and scientific publications in order to detect and track scientific emergence within diverse science and technology communities (Brock et al., 2012; Babko-Malaya et al., 2013a; Thomas et al., 2013; Babko-Malaya et al., 2013b).

Our approach to technology term detection follows from the successful application of supervised learning in information extraction tasks such as named-entity detection (Nadeau and Sekine, 2007) and medical concept extraction from clinical records (Uzuner et al., 2011). The general methodology involves using a large set of human annotated examples of the target class(es) along with their textual contexts to serve as training examples for generating a machine learned model which exploits features extracted from the labeled terms and their contexts. However, unlike the well-defined entity types in those domains (e.g., company names, geographical locations, medical symptoms and treatments), the imprecise definition and immense scope of technical terminology present unique challenges. Consider, for example, the definitions of ”technology” provided by the American Heritage Science Dictionary (Kleinedler and Spitz, 2005):

1. The use of scientific knowledge to solve practical problems, especially in industry and commerce.
2. The specific methods, materials, and devices used to solve practical problems.

The range of terms that fit the second definition above is quite broad, running the gamut from esoteric devices like magnetometers and nanotubes to everyday artifacts like articles of clothing or furniture. Examples from WIPOs International Patent Classification[1], a large multi-level hierarchy designed to support the assignment of patents to categories, follow:

1. Apparatus for the destruction of unwanted vegetation, e.g. weeds (biocides, plant growth regulators)
2. Fittings or trimmings for hats, e.g. hat-bands
3. Geodesic lenses or integrated gratings

For our purposes, then, we define a technology term broadly as a lexical phrase denoting an artifact, process, or field of study (further nuances of this definition are elaborated below).

Since technology development is a global phenomenon, monitoring the life cycle of technologies requires analysts to track literature in many languages. Thus, it is critical that the methodology for technology term extraction generalize readily to multiple languages. To test the generalizability of our approach, we apply and evaluate the methodology on English, German and Chinese patents.

The paper is organized as follows. We first provide an overview of the full system, describing the extraction of candidate technology terms from text, annotation strategy,

[1] http://www.wipo.int/classifications/ipc/en/

generation of training instances, construction of a technology term classifier, and use of the trained model to produce a technology ontology. We then present the results of an evaluation on a subset of English patents, followed by results for German and Chinese. We conclude with a discussion of these findings and opportunities for future work.

2. System Description

New technologies often demand the creation of new sublanguages, while standardization of a vocabulary over time tends to indicate the maturing of a new field. Thus, temporal fluctuations and trends in terminology can assist analysts in their detection and assessment of technology emergence, especially when used in conjunction with other actor-network indicators (Latour et al., 2010). Our goal is the construction of a comprehensive and extensible lexical ontology of technical terms that can serve the needs of text-based analytical tools across multiple languages.

Given the vast number of artifacts and processes described in patents, we opted for a supervised machine learning approach to technical term detection. The feasibility of this approach depends upon both the existence of discriminative contextual features and sufficient training data to enable appropriate feature weights to be learned from examples. To simplify the task, we preprocessed the text using shallow linguistic processing rules to select candidate words and noun phrases; then supervised machine learning was employed to classify these candidates as technology terms or not. The diagram in Figure 1 presents the overall architecture of the system.

2.1. Pre-processing and candidate selection

The patent data used for building the system consisted of small collections of xml-formatted patents randomly selected from LexisNexis' English, German, and Chinese patent databases. Each subset contained 500 documents and spanned the years between 1980 and 2012. Each patent was parsed with respect to its xml document structure to identify relevant sections (title, abstract, first claim, background, etcetera). Then the Stanford tagger[2] was run over the text to detect sentence boundaries, extract tokens (a task requiring word segmentation in Chinese) and assign each token a part-of-speech tag.

Next, a language-specific chunker was used to scan token sequences greedily for the longest sequences matching simple noun phrase patterns. In English, most candidate phrases are of the form (ADJ? N* N). Each part-of-speech tag in a pattern may have an associated list of noise words that are to be excluded from the matched patterns. These serve primarily to eliminate many non-substantive modifiers from the greedy phrase matcher. For example, the leading adjectives ”first”, ”specific”, or ”following” would be considered noise words and excluded from any matching candidate phrase while substantive adjective modifiers like electronic or radioactive would be retained. The output of the chunker is a list of candidate noun phrases along with associated sets of contextual features (e.g., surrounding words and n-grams) which serve as features for machine learning. Similar chunking rules perform the equivalent function in German and Chinese.

2.2. Manual annotation of terms

Supervised learning requires a gold set of manually annotated instances that label terms according to a set of predefined classification criteria. For the purposes of annotating technologies, we defined a technology term as a phrase matching any of the following criteria:

- Artifact – a man-made object produced as the result of a scientific manufacturing process (e.g., electron microscope, computer keyboard)
- Process/technique – the name of a method or process for creating an artifact or doing technical work (e.g., duty cycle control, electron microscopy)
- Field – the name of a discipline or scientific area relating to the production of artifacts or processing (e.g., biotechnology, construction engineering)

In some cases, interpreting phrases using these criteria alone proved problematic. For example, many natural kinds are produced by artificial means, such as smooth muscle cells produced by cell culture or an amino acid sequence determined by protein sequencing. In the context of patents, these typically function as artifacts and hence technology terms. There are some candidate noun phrases which include appositive terms, as in ”clock pulse CK” or ”clock pulse cp1”. Since ”CK” is a generic way to abbreviate ”clock pulse”, the former phrase was considered a technology term whereas the latter, referring to an instance within the patent, was not. A patent typically makes many references to components of an artifact, as in ”resist-free back side”, ”rear cross frame member”, and ”parent identifier field”. Unless these terms refer to components that can reasonably be thought of as independent artifacts, they were not to be considered as denoting technology terms. Also problematic are broad terms which may refer to a technology but in an underspecified manner, such as data or circuits.

In order to reduce the effort required for manual annotation and to maximize its effectiveness for training, we made the simplifying assumption that each phrase (i.e., term ”type”) need only be labeled once, even though some phrase instances might serve different functions in different patents. This simplification relieved the annotator of labeling multiple instances of the same term, a task which would have required considerable work, inspecting each context in which each term appeared within each patent. Instead, the annotator labeled each term within the broader ”context” of technology patents as a whole, deciding based on his/her understanding of a term whether a use of the term would most likely denote a technology. Assigning a label often required the annotator to do a web search to understand the meaning of unfamiliar candidate phrases. (A search for the quoted phrase, sometimes ANDed with the term ”technology” or ”definition” or both, usually produced enough information in the result set snippets to make a decision.) This approach to constructing a training set is a form of ”distant supervision” (Mintz et al, 2009) and runs the risk of introducing

[2] http://nlp.stanford.edu/software/tagger.shtml

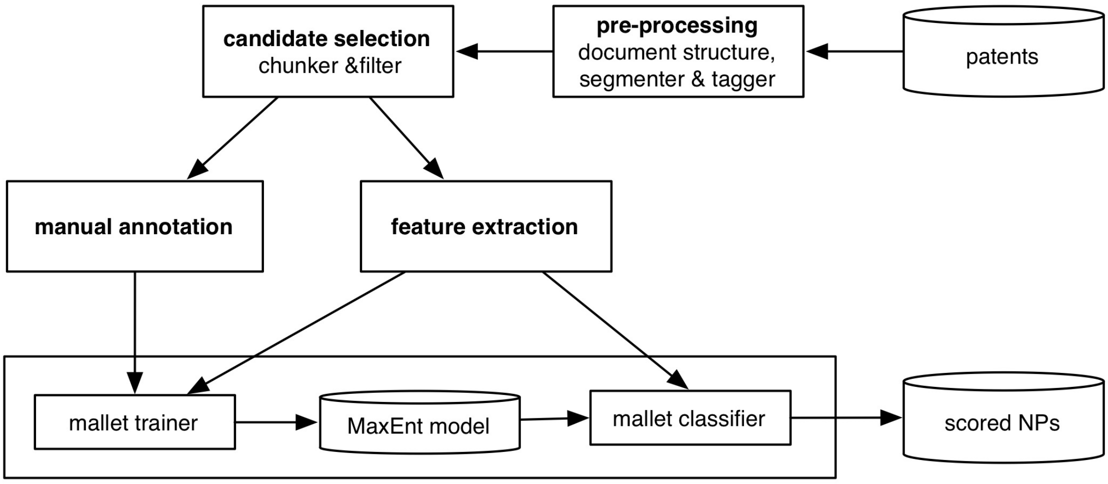

Figure 1: System Diagram

noise. For example, some terms, such as generic single word terms that have several distinct meanings or phrases that may refer to both a natural kind and an artifact, are particularly difficult to classify and indeed may not have a single dominant interpretation in the corpus. Rather than force a decision, we gave the annotator the additional option of labeling a term "?" whenever the annotator lacked the confidence to choose a single classification for the term out of context. Such labeled terms were not included in the gold set for training the model.

Candidate terms for annotation were generated using the output of the chunker and sorted by document frequency so that more common terms were labeled first. More frequently occurring terms would be expected to generate more training instances when applied to the corpus. For each language, annotators provided a minimum of 2000 labeled terms, for English, extra terms were annotated, resulting in a set of 3784 labeled terms. The overall agreement between the annotators, using Cohen's Kappa, was 0.52, suggesting moderate agreement. The annotators were not experts in the technical areas of the patents.

2.3. Features

To create training instances from the labeled terms, each term and label were combined with a contextual features associated with occurrences of the term found within the document collection. Features fell into the following categories:

- External local context: ngrams of size 1, 2, and 3 to the left and right of the term
- External syntactic context: rule-based "dependency" relationships between the term and preceding nouns, verbs and adjectives (prev_V, prev_Npr, prev_Jpr, prev_J). These were intended to capture, for example, the verb (and any prepositions/articles) for which the term is the object. prev_Npr captures a dominating head noun and preposition (e.g., the phrase "a large reduction in the cpu speed" would generate the feature prev_Npr=reduction_in for the term "cpu speed, whereas the ngram context would create the features prev_n1=the, prev_n2=in_the, prev_n3=reduction_in_the).
- Internal features: these include number of tokens in the phrase, first_word, last_word, and suffixes of length 3,4, and 5 characters.
- Document location features: term's location within the structure of the patent, broken down by "1st sentence" and "later sentence" within title, abstract, summary, description, and first claim.

Table 1 shows the total number of potential training instances produced for the 500-document collections in three languages, as well as the percentages of them covered by the most frequent N labeled types. The numbers suggest that a relatively minor annotation effort can generate a significant number of training instances. We will discuss the number of positive and negative examples again in a later section.

	instances	100	1000	2000	5000
English	237,960	10%	29%	36%	48%
Chinese	133,921	21%	49%	60%	75%
German	87,469	20%	50%	61%	77%

Table 1: Share of N most frequent candidate terms

Since the same term can appear multiple times within a single document, there are several approaches to generating training instances for a classifier. We could treat each single term occurrence as a separate instance for training or else merge features from multiple occurrences within a single patent into a single feature vector. While we plan to compare both approaches in future work, for this study we opted for the latter approach, as it allows for a model to be trained directly on the conjunction of features found within each document. Multiple occurrences of the same feature were collapsed into a single feature, rather than counted or weighted. The output of this step, then, was a list of binary feature vectors, one for each term (type) within a document.

2.4. Classification

We used the training data from each language collection to train a maximum entropy classifier using the mallet tool kit (McCallum, 2002). The resulting models can be applied to our task in two different ways. A model can be used dynamically to detect technology terms in a new unseen patent. Alternatively, a model can be applied in batch mode to a large collection to create a global ontology of technology terms. In this mode, the category scores for the same term across multiple documents are merged into a single statistic (e.g., by computing their average, min or max scores). This approach allows scoring for each term to be based on a larger sample of patents, which may lead to more reliable categorization. Building a global ontology off-line also allows for terminology detection in new patents to be done simply and efficiently using dictionary lookup. However, this approach risks lower recall as the global ontology lacks knowledge of any previously unseen terms. A hybrid approach, in which classification scores are dynamically computed for all candidate terms in a new document while global ontology scores are used to bias decisions about previously seen terms may offer the best solution by combining local (document) and global (collection) information. Since the mallet classifier output includes probability scores for each class, it is possible to set arbitrary thresholds for accepting technology terms based on desired levels of precision and recall.

3. Results and Discussion

To evaluate our system, we divided a randomly selected 500-document English collection into a training set of 490 patents and a test set of the remaining 10 patents. Over 3700 candidate phrases from the training collection and nearly 1500 from the test set were annotated with ”y” or ”n” labels. Any terms appearing in the test (”gold”) set were subsequently removed from the training set so that the two labeled term sets were disjoint. A maximum entropy classifier was trained on labeled instances from the training collection. The model thus created (named Model M_1) was used to generate probability scores for the test set terms. Using the gold set labels, precision, recall and f-score were computed for the system-generated results at the acceptance threshold of 0.5. The results are shown below in Table 2.

	P	R	F1
M_1	0.63	0.57	0.60

Table 2: Precision, recall and f-score

We examined high and low scoring terms within the evaluation set to better understand the nature of the false positives and false negatives (Table 3). Among the highest system scoring terms for which the manual (gold) annotation was negative we find some generic artifact terms (”device”, ”identifier”) which may, under the circumstances, have qualified as artifacts. This exemplifies the difficulty of annotating terms for the purpose of classifying artifacts. There is a large class of highly specialized unambiguous terms (such as the true positives shown in the table). At the same time, there is a large class of common terms for which the ”correct label” is less well-defined. To some extent, these terms are not particularly interesting, given that analysts will be interested only in the specialized terms, not the general ones. However, labeled general terms in the training data (and in the evaluation) will impact both the actual performance (and evaluation) of the system. Similar issues arise for some of the negatively labeled terms: ”storage system unit” and ”long extended conductor device” are arguably ”descriptions” of artifacts rather than terms directly denoting artifacts, but nonetheless the labels used for training purposes could have a direct impact on the effectiveness of training data, given that the contextual features for artifact descriptions are likely to be the same as for artifact terms. This suggests a need for further refinement of our annotation guidelines, particularly concerning the proper labeling of generic terms and descriptive phrases.

Low scoring terms with positive gold labels (false negatives) include many single word terms that are unambiguously artifacts: ”database”, ”cpu” and ”solvents”. While it is possible that their roles in the particular patents used for evaluation may have been minor enough to lack sufficient contextual clues to identify them as such, their scores are more likely a symptom related to the class of single word terms.

y	0.988578	graphics processor
y	0.986901	communications system
y	0.986115	computer vision system
y	0.983682	luminescent nanoparticles
y	0.981159	spatial analysis
n	0.933401	long extended conductor device
n	0.892993	coronary artery
n	0.892514	device
n	0.892496	light source
n	0.880899	identifier
n	0.000000	lowered position
n	0.000000	interior
n	0.000000	hook-like part
n	0.000000	highest position
n	0.000000	guide walls
y	0.026642	algorithm
y	0.017968	cpu
y	0.017956	solvents
y	0.017776	pixels
y	0.014474	polymerization

Table 3: High and low scoring terms with their gold labels. Groupings capture true positives, false positives, true negatives, and false negatives, respectively. The table shows the gold label, the system score and the term.

Such observations raised a number of questions about our system design, ranging from the efficacy of specific feature types to the consequences of the distant supervision approach. In particular, we were interested in the following questions:

- Since we are using a large set of labeled ”seed terms” to create training instances through distant supervision rather than annotating each term in context, how is performance affected by the mix of tokens and types appearing in the generated training instances? As the size of the training instance set generated from the seed terms grows, more frequently occurring labeled terms may gain greater representation in the training set. However, the most frequently occurring terms are also the terms most likely to have ambiguous interpretations, which could introduce noise into the training data. Would there be any benefit to setting thresholds for the contributions of frequent types when building the training data?

- What is the relative importance of external contextual features vs. internal information about the term itself (e.g., head word and suffix features)?

- Given the apparent importance of term internal information (head words and suffixes) for classifying phrases and the fact that the vast majority of terms are multiword phrases, how are single word terms (that lack these clues) impacted? Would it be more appropriate to train separate models for single words and phrases?

- Training instances are constructed by joining in a single vector all features related to all occurrences of a term within a document. Would there be an advantage to weighting the feature vector by feature occurrence counts, vs. treating it as a binary (presence/absense) vector?

- Are a term's locations within a patent related to its likelihood to be an artifact? What is the contribution of including location information as features?

- Are the n-gram features preceding the term redundant with or more or less important than the dependency based features? Do both sets of features make independent contributions to the performance?

We conducted experiments to investigate some of these questions. Regarding the issue of transfer of labeled terms from one patent collection to another, we had focused our annotation effort on labeling the most frequent terms in our source collection in order to maximize transfer. However, patents contain many rare and specialized terms and a significant overlap of terms from one set to another, especially across domains, is not guaranteed. To test the effect of training using a set of patents different from those from which our original annotations were drawn, we randomly assembled a different collection of 500 patents, generated training instances from it and tested the resulting model on our evaluation data. The original model M_1 had 3,808 positive instances and 40,589 negative instance, distributed over 1,949 positive types and 1,778 negative types. Building the new model M_2 resulted in 2,880 positive instances and 37,480 negative instance, distributed over 389 positive types and 1,070 negative types. The results are shown in Table 4. As expected, there is a drop in performance, due, most likely, to the decrease in the number of training types generated from this collection.

	P	R	F1
M_1	0.63	0.57	0.60
M_2	0.59	0.55	0.57

Table 4: Precision, recall and f-score for two models of the same size

In an attempt to overcome the performance deficit, we experimented with enlarging the patent collections used as a source of training instances, noting the number of term tokens and types that appeared in the training data as the source collection size was increased. This resulted in a new model M_3 with an optimal size of 10,000 documents, which yielded 58,306 positive instances and 755,156 negative instances, distributed over 689 positive types and 1,437 negative types (which is still significantly fewer than in our original model). Table 5 shows that the larger model does not help increase the precision over the smaller models M_1 and M_2, but that recall increases significantly. Creating models over 20,000 and 50,000 patents showed no increase in precision or recall.

	P	R	F1
M_1	0.63	0.57	0.60
M_3	0.57	0.77	0.65

Table 5: Increasing the size of the model

We hypothesized that the large numbers of instances associated with a few frequent terms may adversely effect the results, especially for those cases where it is not very clear whether a term is a technology or not. To investigate this, we performed two experiments: (1) revising the training gold data of labeled terms and throwing out some of the more unclear frequent terms, and (2) taking a much larger training set of over 350,000 patents and down sample the number of instances per term to a maximum of 1000. The first experiment showed some promise with small training sets, but the effects tailed off for larger training sets and there was no configuration that displayed the same performance as Model M_3. The second experiment resulted in a slightly higher F-score of 0.66.
To gauge the contribution of internal and external features we took the instances as used for model M_3 and built models with only internal features (M_4) and only external features (M_5). Table 6 shows that the overall results are dominated by internal features. Using external features gives a high precision but an extremely low recall. This seems to suggest that technologies in general are not characterized by their linguistic context.

	P	R	F1
M_3	0.57	0.77	0.65
M_4	0.55	0.77	0.64
M_5	0.73	0.04	0.08

Table 6: Internal and external features

We also looked at the impact on the f-score when removing each of the features individually. Most features, when taken out in isolation, did not have much impact on the score. The most notable exceptions was the `last_word` feature, whose removal reduced the f-score by 0.09. The phrase length feature `plen` and the `suffix4` feature both reduced the f-score by 0.02. Note that these are all internal features.

The difference in performance between single-token terms and multi-token terms is shown in Table 7 below. The system labels were created with model M_3, but evaluation was partitioned according to the single-token versus multi-token distinction.

	P	R	F1
all terms	0.41	0.69	0.52
single-token terms	0.20	0.08	0.09
multi-token terms	0.42	0.80	0.55

Table 7: Performance on single-token terms and multi-token terms

Note that the numbers in the ”all terms” row are not the same as the numbers for model M_3 as reported before. This is because the basic evaluation set was too small to allow for meaningful metrics for the single-token terms. We increased the size of the evaluation set, but have not yet performed quality control on this new set. Initial inspection showed a larger percentage of annotation errors that in the basic set, which is probably the reason that precision and recall are lower.

What jumps out is the very low recall for single-token terms. We have not yet determined what exactly is at the core of this.

Comparing the results for classifiers trained on different training sets, we note that precision is highest when the coverage of different terms (types) in the training data is highest (Table 2). Recall appears to benefit more than precision from training sets which include more instances of the same terms. These additional instances provide new contextual features which increase opportunities for generalization. However, the bulk of these additional contexts may be coming from a relatively small set of common patent terms. If even a small number of these common terms are labeled incorrectly in the gold data (or else have multiple interpretations and should not have been assigned a y/n label), these could have an increasingly negative effect as the number of training instances containing them grows. This may account for the slight dip in precision for the larger training set sizes. One way to correct for this might be to limit the number of instances used for any one term so that the contribution to feature weights in the learned model is spread more evenly among different labeled terms.

The growth rate of instances relative to term types as the number of documents in the training set increases suggests that getting sufficient coverage of rare terms in the training data may require very large document sets. Nevertheless, the precision/recall performance for the initial training set, which contains instances of 1033 positive terms and 1407 negative terms, is very encouraging and suggests that increasing the coverage of rare terms in the training set could lead to further improvements in performance.

4. Multilingual Processing

The overall process was essentially the same for Chinese and German, although each language presented several problems of its own. The document structure parser needed some language-specific declarations to deal with useful section headers in Chinese like *technical field* and *background art*. German patents on the other hand had little overt document structure.

Because Chinese does not separate its words using white space, a word segmentation step was required prior to part-of-speech tagging. This was accomplished using a Chinese word segmenter included with the Stanford University language processing toolkit. We used this same toolkit for sentence splitting and part-of-speech tagging for all languages. Patterns for chunking tagged words into candidate phrases had to be constructed for each language. Most contextual feature definitions were sharable among the three languages, with small variations due to syntactic differences. The main time investment in moving to Chinese or German was in the manual annotation. For comparison, we annotated 2000 terms in all three languages.

Abstracting away from the effort to add a segmenter, the time efforts to add Chinese and German versions of the language-specific components were very similar. In both cases it took a computational linguist about a week to adapt the document structure component, integrate the part-of-speech tagger, write chunker rules, define and adapt feature extraction rules and manually annotate terms. An additional day was needed to prepare the evaluation gold standard.

4.1. Multilingual Evaluation

Manual annotation occurred in two phases. In a first phase, which was done for English, Chinese and German, we took the 2000 most frequent technology candidate terms from a training set and associated these manually with ’y’ and ’n’ labels. There was some revision of guidelines and re-annotation, but the focus was on quickly generating labeled instances. In a second phase, which we did for English only, annotation guidelines were given a closer look and a new label ’?’ was introduced which allowed annotators to mark terms that should not be used to generate positive or negative instances. Consequently, the English annotation was completely revised. In addition, extra terms were added to the English term list. In this section, we compare an older version of the English system to the Chinese and German systems, hence, the English results do not match those reported earlier in the paper. The multilingual results are presented in Table 8.

	P	R	F1
English	0.67	0.44	0.53
Chinese	0.52	0.21	0.30
German	0.85	0.36	0.56

Table 8: Precision, recall and f-score for ENglish, Chinese and German

The Chinese system has better precision than the English system at the higher MaxEnt thresholds (not pictured in the table), but recall and f-score lag English scores consistently by a large margin. The lower recall may partially be attributable to a lower number of positive training instances (1286 versus 2496). The German system however has access to a similar number of positive labels as the Chinese system, yet has recall at the level of the English system. We have not yet explained this anomaly. Even more remarkable is the extremely high precision of the German system. This is most likely at least in part the result of a statistical fluke. The German evaluation set turned out to have many less terms than the English one (552 versus 1436) and he numbers in Table 8 are based on small numbers of true and false positives.

The generally lower number of positive and negative training samples for Chinese and German can be explained by the size of the datasets. The 500 English patents comprise 3.7 million tokens whereas the 500 Chinese and 500 German patents contain 1.7 million and 1.3 million tokens respectively.

5. Conclusions

The identification of technology terms within a collection of patents is a challenging information extraction task due to the nature of technology terms themselves, which may be ambiguous or generic and have multiple nuances of interpretation. Initial results using a supervised learning approach are nonetheless very promising and appear to be readily extensible to multiple languages. Our study points to a number of areas for future work, including further refinements to our annotation guidelines and annotation strategy, a better understanding of the relative contributions of additional training terms vs. additional term instances, and the development of strategies for combining term scores from multiple documents. We also plan to compare alternative approaches for the construction of training instances.

6. Acknowledgements

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D11PC20154. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

7. References

Babko-Malaya, O., Meyers, A., Pustejovsky, J., and Verhagen, M. (2013a). Modeling debate within a scientific community. *International Conference on Social Intelligence and Technology (SOCIETY)*, 0:57–63.

Babko-Malaya, O., Thomas, P., Hunter, D., Meyers, A., Pustejovsky, J., Verhagen, M., and Amis, G. (2013b). Characterizing communities of practice in emerging science and technology fields. In *International Conference on Social Intelligence and Technology 2013 (SOCIETY2013)*, State College, Pennsylvania.

Brock, D. C., Babko-Malaya, O., Pustejovsky, J., Thomas, P., Stromsten, S., and Barlos, F. (2012). Applied actant-network theory: Toward the automated detection of technoscientific emergence from full-text publications and patents. In *AAAI Fall Symposium: Social Networks and Social Contagion*, volume FS-12-08 of *AAAI Technical Report*. AAAI.

Fall, C. J., Benzineb, K., Guyot, J., Törcsvári, A., and Fiévet, P. (2003). Computer-assisted categorization of patent documents in the international patent classification (icic'03). In *Proceedings of the International Chemical Information Conference*, Nimes.

Indukuri, K., Ambekar, A., and Sureka, A. (2007). Similarity analysis of patent claims using natural language processing techniques. In *Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on*, volume 4, pages 169 –175.

Kleinedler, S. and Spitz, S., editors (2005). *The American Heritage Science Dictionary*. Houghton Mifflin Company.

Larsen, P. O. and Ins, M. v. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603.

Latour, B., Actant, Callon, M., Law, J., Aramis, o. t. L. o. T., Mol, A., and Verran, H. (2010). *Actor-Network Theory*. Books LLC.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.

PICMET (2012). *Proceedings of PICMET 2012, Technology Management for Emerging Technologies*. PICMET.

Sharma, P., Gupta, B., and Kumar, S. (2002). Application of growth models to science and technology literature in research specialities. *DESIDOC Bulletin of Information Technology*, 22(2):17–25.

Thomas, P., Babko-Malaya, O., Hunter, D., Meyers, A., and Verhagen, M. (2013). Identifying emerging research fields with practical applications via analysis of scientific and technical documents. In *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference (ISSI 2013)*.

Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216 – 1247. ¡ce:title¿Patent Processing¡/ce:title¿.

Uzuner, O., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *JAMIA*, 18(5):552–556.